Meteorol. Appl. 9, 455–459 (2002) DOI:10.1017/S1350482702004085

On the relationship between Clayton's skill score and expected value for forecasts of binary events

Matthew S. Wandishin, *Institute of Atmospheric Physics, University of Arizona, Tucson, Arizona, USA*
Harold E. Brooks, *NOAA/National Severe Storms Laboratory, Norman, Oklahoma, USA*

Following upon previous efforts to associate standard measures of forecast skill to relative economic value, the connection is established between the Clayton skill score and the range of users who realise positive value from the forecasts. It is also shown that, whereas the maximum relative value that can be obtained from a set of forecasts is based on the probabilities conditioned on the observations, the range of users for whom the forecasts provide positive relative value is based on the probabilities conditioned on the forecasts.

1. Introduction

Traditionally, forecast evaluation has focused on measures of quality; recently, more attention has been paid to measures of forecast value. The value approach directly involves the forecast user and the decisions he or she must make given the forecast information. Frequently, this is done in the framework of the cost-loss problem (Thompson & Brier 1955), in which the forecast value is expressed as a function of a forecast user's cost of taking preventative action to the loss incurred when no action is taken and adverse weather occurs. The cost-loss problem is a special case of the more general 2×2 decision problem, in which protection against a loss is not perfect (e.g. Roebber & Bosart 1996). Richardson (2000) has shown that, for the complete range of users having simple 2×2 decision problems, the maximum value of a forecast (relative to the value of climatological forecasts) is given by Peirce's skill score (Peirce 1884), thus showing a connection between the traditional and more recent approaches to forecast evaluation. This paper will demonstrate a parallel equivalence between the range of users for whom the forecast provides positive relative value and Clayton's skill score (Clayton 1934).

2. Derivation

The basic framework of the forecast and decision problem for the 2×2 case is illustrated in Tables 1 and 2, showing the decision and forecast verification problems, respectively, where we assume that all of the numbers can take on non-zero values. (Note that these tables are defined differently in Richardson (2000).) Frequently, the decision process is defined negatively in terms of costs and losses (as described in the preceding paragraph). However, this description is a subset of the more general utility approach (e.g. benefit is obtained by taking preventative action when an event occurs and loss is incurred by taking preventative action when an event does not occur) that will be employed in this paper. The results of the following analysis are independent of the description used.

Table 1. *The 2×2 decision problem expressed in terms of utility.*

		Events	
		Yes	No
Protect	Yes	*A*	*B*
	No	*C*	*D*

Table 2. *The 2×2 forecast problem expressed in terms of joint and marginal relative frequencies.*

		Observed		
		Yes	No	
Forecast	Yes	a	b	$a+b=f$
	No	c	d	$c+d=1-f$
		$a+c=p$	$b+d=1-p$	1

A decision maker in the 2×2 problem is faced with the decision of whether to take preventative action against an adverse weather event given a particular forecast. To illustrate, consider a child's roadside lemonade-stall in a neighbourhood where 20 cups can be sold if the temperature exceeds 30 °C, but only 5 cups can be sold otherwise. Thus, the forecast event is whether temperatures will fail to exceed 30° C. From Table 1, A represents profit realised when preventative action was taken (only 5 cups of lemonade were prepared) and the

event occurred. In this scenario, A will be greater than B, in which the profit is tempered by the lost revenue that could have been realised had the additional 15 cups of lemonade been prepared. Similarly, the utility of C will be equal to the revenue of 5 cups sold minus the cost of preparing the additional 15 cups of unsold lemonade. Maximum utility is achieved in scenario D in which a full 20 cups are sold. The exact relationship between the four utilities will depend upon user-specific details – in the case presented here, the cost of preparing each cup of lemonade and the revenue realised from each sale.

The potential value derived from a forecast involves the interaction between the quality of the forecast and the utility distribution of the user. A 2×2 contingency table relating the forecasts and observed weather, similar to the decision table, can be developed as well (Table 2). At the extremes, if a user always protected the expected utility would be

$$U_{protect} = pA + (1 - p)B, \tag{1}$$

while the expected utility achieved by never protecting would be

$$U_{no\ protect} = pC + (1 - p)D, \tag{2}$$

where $p = a + c$ is the climatological frequency of the event. From these, one can determine that the expected utility of climatological forecasts is given by

$$U_{clim} = max\{pA + (1 - p)B, pC + (1 - p)D\} \tag{3}$$

where *max*{} indicates that we want to take the maximum of the two values. The decision of whether to always protect or never protect depends on the climatological frequency of the event (or base rate), α; specifically, protection should be taken if

$$\mathrm{p} > \frac{D - B}{(D - B) + (A - C)} = \alpha. \tag{4}$$

Provided with perfect forecasts, the user will always protect when the event occurs and never protect when it does not occur, and so the expected utility of the forecasts is given by

$$U_{perf} = pA + (1 - p)D. \tag{5}$$

More generally, the expected utility of forecasts is given by the weighted sum of the probability of the utility associated with that forecast/event combination,

$$U_{fore} = aA + bB + cC + dD. \tag{6}$$

We can set up a standard skill score expression (Wilks 1995) for the potential value of the forecasts with respect to climatology:

$$\mathrm{V}_{rel} = \frac{U_{fore} - U_{clim}}{U_{perf} - U_{clim}}. \tag{7}$$

Plugging (3), (5), and (6) into (7), we get

$$\mathrm{V}_{rel} = \frac{aA + bB + cC + dD - max\{pA + (1 - p)B, pC + (1 - p)D\}}{pA + (1 - p)D - max\{pA + (1 - p)B, pC + (1 - p)D\}}. \tag{8}$$

2.1 Maximum value and Peirce's skill score

First we will recast Richardson's (2000) derivation connecting the maximum possible relative value and Peirce's skill score. Toward this end, consider the solution to this for each argument in the *max*{} expression separately. Taking the first argument, corresponding to the case where $\mathrm{p} > \alpha$ indicating that a climatological forecast dictates that the user should always protect, we get

$$\mathrm{V}_{rel} = \frac{aA + bB + cC + dD - pA - (1 - p)B}{pA + (1 - p)D - pA + (1 - p)B}. \tag{9}$$

To help simplify this expression we can use some identities from the 2×2 forecast problem (Table 2). Specifically, the probability of detection, the fraction of 'yes' events associated with 'yes' forecasts, is defined as $POD = a/(a + c) = a/p$ and the probability of false detection, the fraction 'no' events associated with 'yes' forecasts, is defined as $POFD = b/(b + d) = b/(1 - p)$. (More commonly, in this context, the names hit-rate, H, and false alarm-rate, F, are used for POD and POFD, respectively. However, this has led to some confusion since both H and F have received other definitions as well (see Wilks (1995: 240–1)), and so H and F will not be used here.) From these, we have $a = pPOD$, $b = (1 - p)POFD$, $c = p(1 - POD)$, and $d = (1 - p)(1 - POFD)$. Substituting these expressions into (9) and regrouping, we have

$$\mathrm{V}_{rel} = \frac{(1 - POFD)(1 - p)(D - B) - p(1 - POD)(A - C)}{(1 - p)(D - B)}. \tag{10}$$

Dividing the numerator and denominator by $(D - B) + (A - C)$ and substituting α into the result gives

$$\mathrm{V}_{rel} = \frac{\alpha(1 - p)(1 - POFD) - (1 - \alpha)p(1 - POD)}{\alpha(1 - p)}. \tag{11}$$

If we evaluate (8) for the second argument of *max*{}, the case where $\mathrm{p} < \alpha$, indicating that a climatological forecast dictates that the user never protects against the hazard, we get

$$\mathrm{V}_{rel} = \frac{(1 - \alpha)p(POD) - \alpha(1 - p)(POFD)}{(1 - \alpha)p}. \tag{12}$$

A plot of V_{rel} as a function of α (Fig. 1) shows that this maximises at $p = \alpha$ (= 0.2). At this point, the maximum value is given by

$$max\{V_{rel}\} = POD - POFD. \quad (13)$$

The right-hand side of (13) is equal to Peirce's skill score, and so the maximum relative value of forecasts for the 2×2 decision problem is given by Peirce's skill score. Richardson (2000) notes that, while the location of the relative maximum value along the ordinate of Fig. 1 is determined by p, the magnitude of $max\{V_{rel}\}$ is independent of p, as can be readily seen from (13).

2.2 Range of positive value and Clayton's skill score

The range of users for which the forecasts have positive value is also of interest. The bounds can be found by setting the numerator of (8) to zero and solving for α. For the first argument of the $max\{\}$ function, we get

$$aA + bB + cC + dD = (a + c)A + (b + d)B$$

or, simplifying

$$c(A - C) = d(D - B). \quad (14)$$

Dividing by $(D - B) + (A - C)$ and substituting α as appropriate, this becomes the lower bound for α, α_{min}

$$\alpha_{min} = c/(c + d) = DFR \quad (15)$$

where *DFR* is the detection failure ratio (Doswell et al. 1990), the fraction of 'yes' events associated with 'no' forecasts. For the other argument in (8), the similar expression for the upper bound for users who find the forecasts valuable, α_{max}, is

$$\alpha_{max} = a/(a + b) = FOH \quad (16)$$

where *FOH* is the frequency of hits (Doswell et al. 1990), the fraction of 'yes' events associated with 'yes' forecasts. Thus the width of the interval of users (expressed as a function of their utility ratio α) who find the forecasts valuable is given by

$$V_{wid} = \alpha_{max} - \alpha_{min} = FOH - DFR. \quad (17)$$

The right-hand side of (17) is Clayton's skill score and thus (perhaps not aware of this himself) Clayton was successful in satisfying one of the properties he identified as desirable in a verification method, namely, the 'ability to ascertain at which point weather forecasts cease to have value' (Murphy 1996: 9; Clayton 1889). Richardson (2000) provided equivalent definitions to (15) and (16), but stopped short of (17) and the connection to Clayton's skill score.

2.3 Extension to probabilistic forecasts

The use of probabilistic forecasts (such as those derived from an ensemble system) presents a more complicated system than the 2×2 forecast problem that has been discussed to this point. However, evaluation of a probabilistic forecast can be divided into a series of 2×2 problems based on the set of possible forecast probability thresholds, p_t. Curves of relative value, such as in Figure 1, can be plotted for each p_t (Figure 2). The curve representing the relative value of the probabilistic system as a whole is simply the envelope of these individual curves, i.e. for each α, an appropriate p_t is chosen so as to maximise the value at that point. As shown by Richardson (2000), this means that the determination of the maximum relative value extends naturally from the deterministic to the probabilistic case, namely,

$$V_{max} = max_{p_t}\{POD - POFD\} = max_{p_t}\{PSS\} = PSS_{max}, \quad (18)$$

where *PSS* is Peirce's skill score and $max_{p_t}\{\}$the maximum over all probability thresholds. The width of the interval of utility ratios for which users derive value from the forecasts can also be extended to the probabilistic case, but not as cleanly. The leftmost and rightmost points of this interval can be determined by applying the approach of (18) to (15) and (16),

$$\alpha_{min} = min_{p_t}\{DFR\} \quad \alpha_{max} = max_{p_t}\{FOH\}, \quad (19)$$

but there is no reason to expect that the DFR will be minimised at the same probability threshold for which *FOH* is maximised (see Figure 2) and so

$$V_{wid} = \alpha_{max} - \alpha_{min} = max_{p_t}\{FOH\} - min_{p_t}\{DFR\} \neq max_{p_t}\{\mathrm{CSS}\}, \quad (20)$$

where *CSS* is Clayton's skill score and the ≠ is used in the general sense of 'not necessarily equal to' and is not meant to imply that the *DFR* and *FOH* will never be optimised for the same probability threshold. Therefore, the endpoints of the interval of positive relative value must be computed separately.

3. Discussion

Murphy (1996) examined the similarities and differences between skill measures derived from the 2×2 forecast problem. For this discussion, it is useful to rewrite (13) and (17) in terms of elements in Table 2:

$$V_{max} = POD - POFD = (ad - bc)/[p(1 - p)] \quad (21)$$

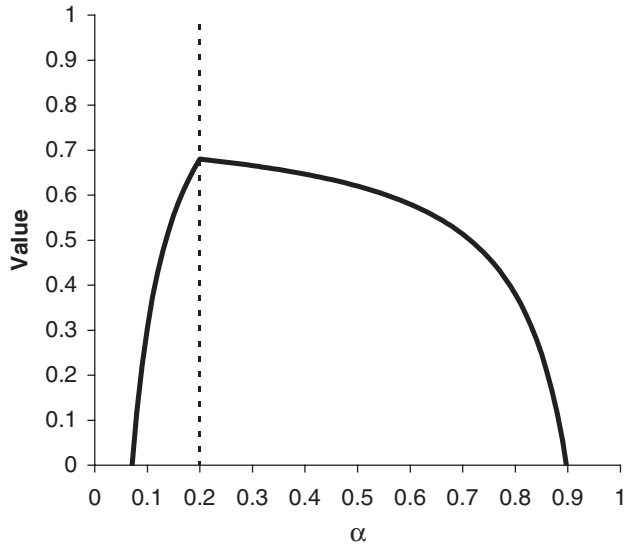

Figure 1. *Example plot of relative value (solid) as a function of utility for a forecast with POD = 0.7 and POFD = 0.02. Dotted line denotes the climatological event frequency (p = 0.2). Note that the value is maximised when the utility is equal to climatological frequency.*

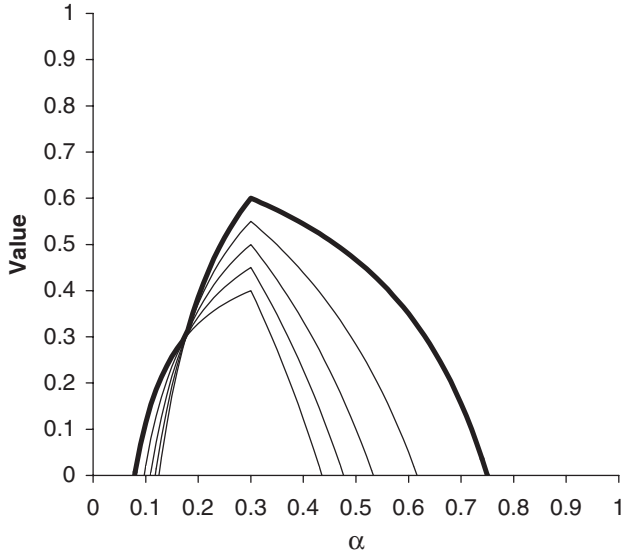

Figure 2. *Example plot of relative value as a function of utility for five different probability thresholds (thin) and the maximum value of the probabilistic forecast system (thick). (POD, POFD) pairs for the five decision points are (0.5, 0.1), (0.6, 0.15), (0.7, 0.2), (0.8, 0.25), and (0.9, 0.3). The climatological frequency is 0.3.*

$$V_{wid} = FOH - DFR = (ad - bc)/[f(1 - f)]. \quad (22)$$

Note that the numerator in each expression is identical and is simply the determinant of the 2×2 matrix in Table 2 (and so positive V_{max} ensures positive V_{wid}, and vice versa). Thus the difference between the two measures is due solely to their denominators: for V_{max} the denominator is the product of the marginal frequencies of occurrence and non-occurrence of the event, while for V_{wid} the denominator is the product of the marginal frequencies of occurrence and non-occurrence of a 'yes' forecast. Thus, V_{max} and V_{wid} can both be expressed in terms of conditional relative frequencies. Specifically, V_{max} measures the difference between the conditional relative frequency that a 'yes' forecast was issued prior to an event occurring (*POD*) and the conditional relative frequency that a 'yes' forecast was issued prior to an event not occurring (*POFD*), the forward looking conditional probabilities (the calibration-refinement factorisation). In contrast, V_{wid} measures the difference between the conditional relative frequency that an event occurred following a 'yes' forecast (*FOH*) and the conditional relative frequency that an event occurred following a 'no' forecast (*DFR*), the backward looking conditional probabilities (the likelihood-baserate factorisation). The denominators in (21) and (22) are in the form of a binomial variance and so high uncertainty (p or f = 0.5) reduces the potential value of a forecast system while rare events (p or f small) present an opportunity for providing more value to more users. In other words, for a given determinant of the 2×2 table (the numerator, $ad - bc$), the maximum relative value depends on the distribution of the observations while the range of users for whom the forecast provides positive value depends on the distribution of the forecasts.

The components of PSS and CSS can also be related to the odds ratio (Stephenson 2000) and thus the odds ratio also provides information on the value of a forecast system. Namely, the requirement for a forecast system to provide value to at least one user, $V_{wid} > 0$ or $FOH > DFR$, could also be expressed as

$$FOH / DFR = (a/b + \theta)/(a/b + 1) > 1, \quad (23)$$

where $\theta = ad/bc$ is the odds ratio. In other words, $\theta > 1$ signifies that the range of users who find value in the forecasts is non-zero. Similarly, instead of computing the difference between *POD* and *POFD* one can look at their ratio,

$$POD / POFD = (a/c + \theta)/(a/c + 1) > 1, \quad (24)$$

and so $V_{max} > 0$ when $\theta > 1$. Therefore, the odds ratio provides information on the existence of positive value but not its extent or the range of users for whom value exists.

It is interesting to note that for unconditionally unbiased forecasts ($p = f$), $V_{max} = V_{wid}$, and so improving the forecasts of an unbiased system will increase equally the maximum value and width of the interval of users who find the forecasts valuable. Also, Murphy pointed out that the product of Peirce's skill score and Clayton's skill score (or $V_{max} * V_{wid}$) is equal to the binary-event version of the Pearson product-moment correlation coefficient (r), and so these scores represent

not only a connection between skill and value, but also between these two aspects of forecast quality (skill and value) and association.

References

Clayton, H. H. (1889) Verification of weather forecasts. *Am. Meteorol. J.* **6:** 211–219.

Clayton, H. H. (1934) Rating weather forecasts. *Bull. Am. Meteorol. Soc.* **15:** 279–283.

Doswell, C. A., Davies-Jones, R. & Keller, D. L. (1990) On summary measures of skill in rare event forecasting based on contingency tables. *Weather and Forecasting* **5:** 576–585.

Murphy, A. H. (1996) The Finley affair: a signal event in the history of forecast verification. *Weather and Forecasting* **11:** 3–20.

Peirce, C. S. (1884) The numerical measure of the success of predictions. *Science* **4:** 453–454.

Richardson, D. S. (2000) Skill and relative economic value of the ECMWF ensemble pediction system. *Q. J. R. Meteorol. Soc.* **126:** 649–667.

Roebber, P. J. & Bosart, L. E. (1996) The complex relationship between forecast skill and forecast value: a real-world analysis. *Weather and Forecasting,* **11:** 544–559.

Stephenson, D. B. (2000) Use of the 'odds ratio' for diagnosing forecast skill. *Weather and Forecasting* **15:** 221–232.

Thompson, J. C. & Brier, G. W. (1955) The economic utility of weather forecasts. *Mon. Wea. Rev.* **83:** 249–254.

Wilks, D. S. (1995) *Statistical Methods in the Atmospheric Sciences.* Academic Press, 467 pp.